DOCUMENT RESUME

ED 413 748 FL 024 769

AUTHOR Teubert, Wolfgang
TITLE Language Resources: The Foundations of a Pan-European Information Society.
PUB DATE 1995-00-00
NOTE 25p.; In: Language Resources for Language Technology: Proceedings of the TELRI (Trans-European Language Resources Infrastructure) European Seminar (1st, Tihany, Hungary, September 15-16, 1995); see FL 024 759.
PUB TYPE Opinion Papers (120) -- Reports - Descriptive (141) -- Speeches/Meeting Papers (150)
EDRS PRICE MF01/PC01 Plus Postage.
DESCRIPTORS *Computational Linguistics; Computer Software; *Computer Software Development; *Discourse Analysis; Foreign Countries; Information Technology; *Language Research; Linguistic Theory; Technological Advancement
IDENTIFIERS Europe; *Language Corpora; Trans European Language Resources Infrastructure

ABSTRACT

Language, once a cultural asset in Europe, has become an economic commodity. The national language institutes now have the task of providing the means for making knowledge available in the national language and distributing locally produced information worldwide. This includes not only training more translators, but also developing the necessary language technology. Current issues in language engineering include development of high-quality, economic document creation and management, creation of better information and communication services, better translation and support of foreign language learning, and internationalization and localization of software for language applications. The quality of language technology applications rests in the comprehensiveness and reliability of the language data with which the tools are used. Corpora are the basic language resources, which exist in a variety of forms, including special, reference, monitor, opportunistic, comparable, and parallel corpora. Building a corpus requires encoding decisions, documentation, and validation; the process is costly. The Trans-European Language Resources Infrastructure (TELRI) is a cooperative effort by language and language technology institutions in 17 countries to supply the natural language processing community with monolingual and multilingual language resources. TELRI will make Europe a strong competitor in the emerging market of language engineering and communication technology. (MSE)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 413 748

Language Resources: The Foundations of a Pan-European Information Society

Wolfgang Teubert

Institut für deutsche Sprache
Postfach 1016 21
D-68016 Mannheim
Tel/Fax: +49 621 1581 415
E-mail: wolfgang.teubert@ids-mannheim.de

PERMISSION TO REPRODUCE AND DISSEMINATE THIS MATERIAL HAS BEEN GRANTED BY

Norbert
Vd3

TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)

This document has been reproduced as received from the person or organization originating it.

Minor changes have been made to improve reproduction quality.

Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

BEST COPY AVAILABLE

FL024769

0. Language resources for language technology: academia meets industry

Language engineering is the core of information technology, and information technology will be the key industry of the 21st century. The information super highways conceived today will soon transport infinite amounts of digital data, images, sounds, tables, figures, calculations, and process protocols. If these data are to be intelligible, if they are to make sense, they must be bound together by language. Without natural language processing (NLP), information remains incomprehensible.

More than any other continent, Europe is multilingual by common commitment. This situation provides a challenge to European language technology. We all want information to cross borders freely; however, countries can only uphold their cultural and linguistic identity if all the relevant information is available and accessible in the national language(s). This is an important principle of the European Union today, and it also holds for all European nations who have not yet joined the European Community. For the emergent European information society, we have to develop a language technology that meets the multilingual challenge. It will have to support the production, revision, conversion, presentation, publication, documentation, and last, but not least, translation of texts in technical and everyday language; and it will have to grant language-independent retrieval by sophisticated interaction modes based on natural languages. This demand calls for a modular solution. Language independent algorithms have to be combined with language specific software. The goal must be to impose as few restraints as possible on one's language in computer-aided communication activities. A "controlled documentation language" designed to avoid all kinds of ambiguities and undesirable complexity will impede the generation and transmission of novel content and concepts. A language without fuzziness, opacity and metaphorisation brings about a petrification of ideas. Only principally unrestricted use of one's language (and the reunification of any "controlled documentation language" in man-machine interaction will guarantee the creativity of information exchange. If this challenge is met, European language engineering will play a leading role on the global market.

The quality of all language technology rests on the linguistic knowledge determining the algorithms of any NLP application. This linguistic knowledge is accessible in and from language resources. We find it in scientifically designed text corpora and in corpus-based or, at least, corpus-validated lexicons; we can extract it from textual and lexical resources by recursive generalisation processes and convert it into the form needed in a specific application. For corpus analysis and lexicon generation, we need powerful generic soft-

ware that will have to be modular to take care of language specific and language independent aspects. It would be erroneous to believe that the linguistic knowledge needed for sophisticated language technology applications is already available. So far, language studies have indeed accumulated huge repositories of data and their interpretation. However, these data were collected for human users only. Human users are able to draw on analogies and can apply inductive reasoning. They have common sense and understand (to some extent) the world they live in. Therefore a human translator cannot be a model for machine translation, nor can bilingual dictionaries be the source for lexicons in translation systems. The task of machine translation (MT) can be compared to the task of translating a text on a remote subject field from an unknown language A into an unknown language B with the help of a bilingual A/B dictionary. A MT lexicon would have to contain all the information to make such a translation possible. The necessary linguistic knowledge has to be gathered from scratch. We do not find it in existing dictionaries. There is no alternative to text corpora.

Corpora are the raw material of all language technology. The better they are, the more expensive is their creation. Language industry, small-and medium-sized enterprises in particular, often cannot afford to build them up. On the other hand, in practically all European countries there are focal language centers with a long tradition in the creation (and also in the application) of language resources. This is not only true for all of Western Europe, but also for most of Central and Eastern Europe including the former Soviet Union. In its short history, computational linguistics always has been a global discipline; and the NLP community was and continues to be well connected.

However, while the results of academic research traveled freely (with a few clandestine pockets here and there), language resources, corpora, and machine readable dictionaries did not flow as easily. Due to the lack of hardware compatibility, the old restrictions now often have given way to new limitations based on property rights. Solutions only can be found in joint efforts of research and industry. We need a European network of all academic research institutions to allow the fee flow of language resources between all partners under fair conditions.

A research project, however, is only the first step. In a second step, we have to set up an operational infrastructure of (public domain) research and (private) industry. We need a common platform where providers and users of language resources come together, share expertise, discuss their needs, exchange resources, join forces, and give birth to new visions. Private industry will ensure that new language technology applications find a market (and the money invested will not be wasted), and public domain research will provide the linguistic expertise to make the products a success.

In some European countries, such a national infrastructure already exists; in others, it is gradually evolving. Still most of the work is devoted to monolingual applications. Until some years ago, there was not much cross-border cooperation at least not in academic circles. This is why repeated efforts in Western Europe have been made to set up a transnational infrastructure that can serve the needs of multilingual language technology applications. In March 1995, the European Language Resources Association (ELRA) was founded with strong backing by the Commission of the European Community. It will be seen in the years to come which additional measures are necessary to secure Europe's role as a leading actor in the global language engineering market. This future Europe will be larger than the European Union of today. Linguistic expertise, language resources, and computational linguistics are highly developed in the countries of the former Soviet Union and in Central and Eastern Europe. We can observe the emergence of a dynamic, if still small, language industry in these locations. If we want to make Europe a competitor on the world market of language industry, we must build up a common infrastructure all the way from Galway at the West Coast of Ireland to Vladivostok at the Sea of Japan.

In most European countries, we find one or more focal academic institutions dedicated to document, analyze, and describe the national language(s) and to put linguistic knowledge across to the public. Language was regarded as the paramount cultural asset responsible for ethnic or national identity. In all research work, the main emphasis was, therefore, on this language; it was important to demonstrate that this language was as well-researched as any other major world language. Communication across borders, from language to language, if encouraged at all, was rarely a topic for these institutions. The transnational flow found its limitations in the number of translators and translations, serving as a useful filter of incoming and outgoing information.

Today, the situation has changed: European integration is now imperative to our societies. A precondition is the uncurtailed accessibility of information. National economies can successfully compete on the global market only if all relevant data are available. However, although it is relatively easy to obtain all the information wanted, it is usually written in a foreign language, unless you happen to speak English or perhaps French as your mother tongue. So, the large majority of people who do not speak a foreign language are put at a disadvantage. German companies regularly complain that calls for tender issued by the Commission of the European Communities are initially and at times only published in French and English: this makes it much more difficult for them to compete. International TV channels first thought that they could reach their audience by transmitting their programmes in English;

however, programme sponsors now insist on subtitles or dubbing in the local languages. Most viewers prefer programmes in their native language even if they understand English.

Language, once a cultural asset, has now become an economic commodity. The focal national language institutes have acquired new responsibilities. Now, their task is to provide the means that information from abroad can be made available in the national language and that locally produced information can be distributed world-wide in major international languages and in the language of the neighboring countries. To train and employ more translators is not sufficient. We have also to take care that the necessary language technology is being developed.

1. Current Issues in Language Engineering

Spelling checkers were among the earliest successful language technology applications. They have been accepted as useful devices and are still being sold today in ever-improved versions. On the other hand, more ambitious projects often have failed. The majority of early machine translation systems, particularly the more sophisticated ones have not survived. SYSTRAN is still kept alive by the European Commission's translation services, but many others have disappeared without leaving a trace.

Spelling checkers do not need semantics. Even in the seventies, they were based on little more than a list of the most frequent word forms, i.e., linguistic knowledge widely available or easy enough to generate from corpora. Machine translation, however, needs semantics. Our understanding of word meanings or lexical semantics in the seventies was contained in dictionaries, and it was arranged in a form that a human user, with some experience, could understand, using a great deal of implicit knowledge about the world and inductive reasoning and having the ability to draw analogies–all assets computers usually do not possess. Therefore, it is no surprise that early applications involving semantics were not very successful.

Today, we know that the semantic data needed for sophisticated language technology cannot (or only to a very small extent) be derived from dictionaries designed for human users. Rather, it has to be generated from scratch, namely, on the basis of corpora. Language technology can work with two kinds of semantic information. The one is rule-based and presupposes an intellectual semantic analysis of the phenomenon in question. The other kind uses statistics and is not really semantic at all: it computes the five or ten words or word forms preceding or following the word in question and relates

this information to the different translation equivalents found in parallel corpora. The German word *Schnecke*, e.g., is translated into English either by *snail* or *slug*, where *snail* refers to the creature with a 'house' and *slug* without one. The rule-based approach states just that and searches the German text for clues from which we can infer the correct translation equivalent. The statistic-based approach does not look at the meaning at all. It looks for words and other traces that frequently occur when *Schnecke* is translated as *snail* and for other patterns co-occurring with *Schnecke* being translated as *slug*. In the context of *slug*, we would probably find words like vegetable *(garden)*, *lane*, *wet*, and various forms of *get rid of*; while in the case of *snail*, I would expect words like *table*, *course*, *wine*, but also *vineyard* and *sunny*.

Today's successful applications involving semantics work with an amalgam of the rule- and the statistic-based approach. The statistical approach has some attractive advantages: the data required can be generated form corpora with no or only little human intervention; and since it is just an emulation of semantics, one does not have to be able to state explicitly what a lexical item means. Indeed, it leaves out the entire question of meaning. Its inherent shortcoming is the rate of accuracy. Even a rate of 95% of correct translation equivalents implies that every twentieth word in a text is mistranslated, practically every sentence and certainly more than what most people would like to live with. On the other hand, the rule-based approach can be a very expensive alternative. It presupposes something like a bilingual dictionary that would enable translation of a text correctly into an unknown foreign language. The reason why such dictionaries do not exist for human users or machines is that the explicit linguistic knowledge they would have to contain is not yet available and that this knowledge is be extremely expensive to produce.

The new generation of language technology applications (monolingual and bilingual or multilingual ones) deals with semantic problems. They recognize the fact that computers cannot understand spoken or written texts in the way humans can. Therefore, these text processing systems can only emulate the human faculty of 'understanding' by a mixture of rules and probabilities. To find the right mix is less a question of theory and principles than of calibration and learning by doing. The crucial point is the performance of an application under real life conditions. The application has to prove its cost efficiency, i.e., it must demonstrate that it can complete a task cheaper than a trained human. After two decades of experimental and pilot systems, the emphasis today is on robust applications for which there is a real market, i.e., one where users are willing to pay a fair price.

This new state of the art is reflected in the policy of the European Commission on information technology. Under the Telematics 4th Framework

Programme, only such projects are funded where the usefulness of the application under development is verified and attested by an industrial interest group, representing the end users of the application. In addition, the project consortium must include partners from academic research and industry. Industry can claim only 50% of its development cost for funding. Thus, the companies involved in such a project must hope to gain the other 50% of their financial investment from product sales. Market orientation is a top priority for the Commission. This is reflected in the list of topics for funding mentioned in the Work Programme of the 1994-1998 Telematics Applications Programme. (1) I will present the following topics:

- Document Creation and Management
- Information and Communication Services
- Translation and Foreign Language Acquisition
- Globalization and Localization of Software

1.1 Document Creation and Management

The Work Programme lists the following objectives: *to produce high-quality documents faster and more economically; more effective handling of and localization of product documentation; faster and more accurate drafting in a non-native language; improved checking of style, grammar and comprehensibility; more consistent use of technical language and terminology; better reuse of existing texts; transparent access to in-house and remote document translation services; automatic indexing and categorization of documents; ability to locate and browse documents; content analysis; automated generation of documents from database records.* (2)

The following applications are mentioned in the Work Programme:

- Authoring of office documents
- Collaborative authoring of technical documents
- Document management
- Document interchange and message handling
- Report generation

All of the applications make use of tools that have existed for some time now such as bilingual lexicons, thesauri, terminological databases, text editors, content analysis procedures, knowledge extraction processes etc. Some of them deal only with formal aspects of texts like editors including spelling and syntax checkers; some use rule-based semantics like (monolingual or multilingual) thesauri, particularly in connection with word sense disambiguators. Content analysis and automatic indexing is based on statistic methods. The

applications mentioned bind together a number of tools in order to be able to fulfill a particular task, a task that would otherwise have to be carried out by humans. The goal is to relieve people from tedious routine like repeating ready-made text elements and to give them more time for writing new text segments.

These procedures also can ensure completeness and consistency of highly formalized texts. Minutes of meetings are a typical example for a document type that is strictly formalized in composition and that involves much routine. As has been shown, adequate software can improve the quality of employment references by asking the author to go through exhaustive check lists.

The quality of applications in this field depends on the architecture that holds together the modules. This relates to the linguistic and technical specifications for the data interchange between the various tools contained in an application. The other important quality parameter is the fine-tuning or calibration of the system for the particular task for which it was designed. What is an appropriate relationship between stylistic variation and consistency in the use of terminology? How much choice should the author have in the organization of the document, in the vocabulary, or in the complexity of syntax he is to use? Less freedom will be permitted if the documents are to be machine rather than hand translated.

1.2 Information and Communication Services

The Work Programme lists the following objectives: *better and more selective data access; more intuitive language interfaces; content-based information analysis and filtering; technology-mediated interpersonal communication services with advanced spoken and written language capabilities; content-based information browsing and navigation; maximum localization of human-computer interfaces.* (3)

The following applications are mentioned in the Work Programme:

- Text- and voice-based communication services
- Access to information and transaction services
- Mobile information and communication services
- Interface and software globalization

As we can see from the objectives mentioned, emphasis in this topic is on semantics: 'more intuitive language interfaces', 'content-based information analysis', 'content-based information browsing'. For applications in this field, it will not be sufficient to use existing tools and assemble them as modules in

a new configuration. Of course, there are already many speech recognition systems. While the majority of them are still experimental, some are already being marketed to end users. But they are still far away from being satisfactory. Speech to text software, dictation systems still impose severe restrictions on the user, e.g., making distinct pauses between the words. They need extensive training for each speaker, and their accuracy rate makes post-editing necessary. Better performance can be expected from speech understanding systems that will look for the most plausible reading of an utterance and not just match the phonetic input with a possible orthographic output. But, most of those systems are still in their pilot stages. To make their performance more reliable, we will need more semantic data that can tell us which reading is more plausible than the alternative ones.

More semantic data than currently available are also needed for any kind of content-based information extraction. As I said before, these data should not be taken from existing dictionaries but from corpus analysis. Content-based analysis talks about concepts and the semantic relationship between concepts, not about expressions and the syntactic relationship between expressions. The quality of information extraction rests with the thesaurus and its links of synonymy, hyponymy etc. that it establishes among lexical units. Most existing dictionaries show these links only between single words, but this view is too narrow. One of the results of the computational analysis of language is that the lexical units to be considered often are multiword units, collocations, and phraseologisms. Applications in this field will have to use better lexicons describing such entities if they want to be superior to existing tools. Building corpus-based lexicons, however, is rather expensive. Indeed most existing lexicons have been converted from dictionaries which are not corpus-based. The creation of a new generation of standardized homogeneous generic corpus-based lexicons for the languages of Europe is a major task, and it certainly costs more than small and medium-sized enterprises can afford. This job also demands the kind of linguistic expertise one finds more in academic institutions. Close cooperation between academic research and private industry, therefore, is an important condition for the development of robust applications in the information and communication services.

1.3 Translation and Foreign Language Acquisition

The Work Programme lists the following objectives: *lowering of language barriers; more effective support for professional and occasional translators; access to in-house and remote translation facilities; better support for the translation of*

technical documents; speech translation of business documents and correspondence; more effective learning of lexical and grammatical knowledge; more effective foreign language reading and writing aids. (4)

The following applications are listed in the Work Programme:

- Translator's toolbox and translation aids
- Machine-aided translation
- Tutored and self-trained foreign language acquisition

It is noticeable that the goal of machine translation per se has been given up. The huge EUROTRA project has failed in a number of aspects; machine translation of unrestricted text without extensive post-editing is not possible. Still, machine translation is better than no translation–it gives the reader some idea of what the text is about. Machine translation is possible and cost-effective (in spite of the necessary post-editing) where huge amounts of relatively standardized texts in a specific domain (i.e., airplane maintenance manuals) have to be translated. Its quality can be improved by restricting the input: restrictions on the vocabulary and the syntactic complexity of sentences, on the use of anaphoric pronouns, on coordination structures etc. Thus, it is no surprise that weather forecast messages are a typical example for successful automatic translations from French into English and vice versa.

In most situations, however, it is preferable to develop tools enabling human translators to produce better translations in less time. In this field, again, semantics is the core issue. The meaning of a text should remain constant, regardless of the language used. The task of creating bilingual or multilingual lexicons is enormous. Existing dictionaries are of little help. They are good enough for human translators to render a foreign language text into their native language. However, due to a lack of relevant semantic data, they are not sufficient to translate from one's native language into a foreign language.

The basis of new tools to facilitate human translations should include parallel corpora, i.e., texts and their translations into other languages. They contain, though only in an implicit form, all the linguistic knowledge a translator needs; however, this knowledge has to be extracted from the corpus of parallel texts before it can be reassembled in tools that are the core of novel translation aids. Here again, the lexical unit that is translated is not only the single word, but (more often) multiword units, collocations, and phraseologisms. Bilingual and multilingual lexicons will be repositories of translation equivalents, regardless of whether they are single words or complex phrases extracted from parallel corpora and arranged according to linguist principles. Therefore, the methodology for lexicons of translation equivalents will probably be developed and perfected in academic research. On

the other hand, how much of the information and in which sequence it will be presented to the translator is a matter of fine-tuning and will vary from application to application.

The conception of new translation aids, such as the translation memory, must have a high priority in European language technology. For a long time, academic linguistic research and the variation of language resources have focused on the monolingual situation. However, if information is to travel freely in a multilingual Europe, language barriers have to be lowered considerably by new translation tools. Multinational software houses were the first to take up the challenge of multilingual language technology. Yet, even they lack the specific language resources needed for a robust performance of multilingual applications. Their creation is the most immediate task for all the nations desiring participation in the European information society.

1.4 Globalization and Localization of Software

The field of globalization and localization of software is not a separate chapter in the Work Programme of the new Telematics Applications Programme. However, it is briefly mentioned under the heading 'Information and Communication Services': Multilingual products and services presuppose flexible and effective methods and tools for human-computer interface and *software internationalization and localization, leading to faster and more widespread user acceptance.* (5)

From the point of view of academic institutions and small- and medium-sized private industry in Central and Easter European countries and of the former Soviet Union, the more interesting aspect is globalization of resources, generic tools, and specific applications. There are linguistic data, and there is sophisticated software, but it is not used outside the country or language area where it was designed. Of course, there are language specific tools like lemmatizers and morphological generators for which there seems to be no use outside the particular language; however, a tool like a lexicographer's workbench or, to be even more general, a tool for language data compression can be used for any language. Complex applications usually will be conceived in a modular way. Some modules will be language specific, while others are language independent. The emergent European language resources and language technology infrastructure will provide a platform where academic and industrial partners in numerous countries can be found willing and competent to adapt existing tools to other languages and to distribute such tools in Europe and on a global scale.

This is true not only for tools but also for corpora and lexical resources. Even if they were originally composed for monolingual applications only, it is worthwhile turning them into multilingual resources. Some texts in a corpus might have been translated in other languages; if these translations are included, then there is a new parallel corpus for which there will be demand in other parts of the world as well. There are several ways to establish semantic links between entries of monolingual lexicons. Thus, it is possible to constitute a multilingual lexical database on the basis of existing monolingual lexicons. These lexical data can then be used as input for the alignment of translational equivalents in parallel texts, and this procedure in turn supplies new semantic links between the entries of different languages. Globalization of lexical resources is a precondition for the development of all multilingual applications and should be given high priority. The evolving European infrastructure must encourage joint activities in this field.

Localization of software that has been developed for the English-speaking market is the other face of the coin. Localization is necessary for complex tools designed for nonspecialist end-users, whenever the language barrier would keep them from using it. This is clearly the case for software like text editors, style checkers, or other authoring tools for any software that handles or manipulates language. Every language specific module has to be redesigned for the local language; also, some adjustments usually have to be made for the more language independent modules. For these and for all other tools, which are to be used by a wider audience who does not necessarily understand English, texts used in front ends have to be localized, and the user manuals have to translated. There is a quickly growing demand for localization services; and for academic institutions and small local software houses, it can be very attractive from the financial point of view to offer this kind of service. Contacts with the huge global and predominantly English-speaking market can, thus, be established and used later for other activities as well.

Information technology is constantly undergoing change, change that is reflected in the terminology of this discipline. Institutions working in the localization field have immediate access to new terms as they occur in manuals and other technical product descriptions. Thus, these institutions have a strong voice in the creation of the national terminology. Experience has shown that whoever can exert control over the terminology of a given discipline also often has a leading edge over his or her competitors. This is another reason why it can be attractive to set up a localization service.

One of the three 'colleges' in the field of language resources (besides speech and written language) is terminology. Key industries today are organized as global markets, and their terminologies are strongly dominated by the English

language. In order to sell products on local markets, however, advertisements, product descriptions, maintenance manuals, and other technical documents have to be available in the language(s) of the respective country. In the seventies and the early eighties, some international companies believed that people would buy their products even if they were described only in English; however, there is a lot of evidence that whenever consumers had an alternative they preferred the product that was fully documented in their native language. The localization of international (mostly English) terminology is becoming an important activity in the field of language resources. The European Commission has funded terminology infrastructure projects in Western Europe and also, under the COPERNICUS programme within all of Europe including the Commonwealth of Independent States. (6)

2. The Importance of Language Resources

As we have seen in the preceding chapter, the quality of language technology applications rests foremost with the comprehensiveness and reliability of the language data with which the tools are used. For traditional applications like a spelling checker, 'hard' data are needed: word forms (including orthographic variants) and a morphosyntactic analysis leading to lemmatization. For more sophisticated applications that somehow involve meaning, we need 'soft' data, data which are interpretations of hard data by competent linguists. Finally, for multilingual applications, we need soft data for various languages and procedures allowing the mapping of these data onto each other, again necessitating interpretation by competent linguists.

Hard and soft data together constitute our linguistic knowledge. We have argued that today's knowledge is neither reliable enough nor sufficient for the development of robust heavy duty tools performing noticeably better than the existing toy or pilot systems.

Many computer linguists, particularly those with an engineering background, take it for granted that, with the linguistic knowledge available today, a steady improvement of language technology systems is possible to the point where, say, an operational, robust system for machine translation can be developed. They are impressed by the fact that seemingly astonishing results can be achieved with stochastic methods necessitating little knowledge and even less interpretation of soft data of the language(s) involved. They believe to be justified in their convictions because they look at natural languages as nothing but a set of particularly complex formal languages. However, there is a generic difference. Given basic conditions, formal languages

can be translated into each other. But anyone who has translated from one natural language to another knows it takes more than just the hard linguistic data plus a few statistic operations. Translators must also understand the content of the text and must be aware that they will understand it only if they have sufficient knowledge of the world. This is not knowledge about facts rather now we interpret the objective world that we perceive. To some extent, this (community agreed) interpretation will have to be integrated into the language data.

As I stated before, the linguistic knowledge available--the knowledge formulated in existing grammars and dictionaries--is unsuitable for language technology tools for two reasons. First, most of it is not corpus-based; rather it reflects the individual linguist's (lexicographer's) competence based on a collection of data (citations); and, however large their collection may be, it is permeated by a bias that cannot be avoided. Second, traditional grammars and dictionaries have been devised for human users, and human users differ substantially from machines. Humans use inductive reasoning and can draw analogies easily; faculties like these are taken for granted and are reflected in the traditional arrangement and presentation of processed linguistic data. Language technology tools cannot take recourse to common sense. for language technology applications, all knowledge has to be spelled out in the form of rules, lists, and probabilities.

This task is sufficiently demanding in itself. In order to carry it out, we have to go back to the sources; and the only source for linguistic data is the corpus, the authentic and actual texts in their unannotated representation. But, anyone who has gone to the sources has also experienced the problem that when we start analyzing language as it occurs in a corpus, we gain evidence that renders existing grammars and dictionaries as very unreliable repositories of linguistic knowledge. We discover that our traditional linguistic knowledge gives us a very biased view of language, a view that has its roots in the contingency of over two thousand years of linguistic theorizing. We are so accustomed to this view that we take it for the truth, for reality, not just for an interpretation of hard data. It is true that traditional grammars and dictionaries have helped us, fairly satisfactorily, to overcome the linguistic problems we humans have to deal with. But they will not suffice for language technology applications.

This is why, however cumbersome and expensive it may be, language has to be described in a way that will be appropriate for language engineering. In the Council of Europe corpus project - the *Multilingual Dictionary Experiment* (project leader, John Sinclair, with participants from Croatia, England, Germany, Hungary, Italy, Sweden) - it has been demonstrated that

monolingual and bilingual dictionaries are of no (or only little) use when it comes to automatically translating a word from one language into another in cases where there is more than one alternative. (7) A close analysis of the problems involved in the translation of nominalizations between German, French, and Hungarian (also corpus-based) has shown that all the descriptions available in traditional dictionaries and grammars are inadequate, incomplete, and ultimately useless. (8) To reduce the cost of a corpus-based language analysis from scratch, which is indispensable, corpus exploitation tools have to be developed which arrange the hard facts (including statistic-driven devices for context analysis) and which process them (with a great deal of human intervention for the semantic interpretation of data) into algorithmic linguistic knowledge, into rules derived from objective data rather than individual competence. Perhaps, this will result in the finding that traditional categories like noun, verb, and adjective do not, after all, reflect categories useful for NLP.

Corpora are the basic language resources. They have been used by linguists for about thirty years. We have learned that the early discussion on the representativeness of corpora led in the wrong direction. Corpora represent nothing but the texts they consist of, and certainly not a language universe. The analysis of corpora has given rise to new insights, particularly concerning the vocabulary. In the sixties, we were accustomed to think that we could define the core of general language as the intersection of special languages and that such a definition would make it possible to define a finite general language vocabulary of perhaps 50 000 to 100 000 lemmata. Today, we are much less sure about the usefulness of such constructs. Instead, we talk about balanced corpora, corpora composed according to parameters like text type (e.g., 3rd person, present tense only), genre (e.g., romance, instruction, free asymmetrical conversation) and domain (e.g., angling, crime, stock exchange).

We know that a balanced corpus regardless of size, number of parameters, and numbers of values assigned to each parameter does not represent general language; instead, we say it can be used for a number of different purposes. It does not represent the vocabulary of general language because a general language vocabulary is not a meaningful concept. All we can say is that we aim at a corpus that is 'saturated' in terms of the vocabulary. This means that a particular chunk of our balanced corpus representing one text type, one genre, and one domain (e.g., texts in the 1st and 3rd person, past, and present tense; newspaper diary; angling) is saturated once the growth rate of the vocabulary stops decreasing and becomes constant. There is no point from which there will be no hitherto unrecorded words, but there is a point from which there will be perhaps eight new words (types) for each 10 000 ad-

ditional words of text (tokens). Saturation of corpora is a fairly new concept, and no one knows yet what it leads to in terms of corpus size.

John Sinclair has recently developed a corpus typology, and I am using it in my brief account of corpus types. (9)

Special corpus. Special corpora are assembled for a specific purpose, and they vary in size and composition according to their purpose. By intention, special corpora are not balanced (except within the scope of their given purpose) and, if used for other purposes, give a distorted view of the language segment. They can have a number of advantages compared with balanced corpora. Their main advantage is that the texts can be selected in such a way that the phenomena one is looking for occur much more frequently in special corpora than in a balanced corpus. A corpus that is enriched in such a way can be much smaller (perhaps ten times) than a balanced corpus providing the same data.

Reference corpus. Reference corpora come closest to the old concept of a representative corpus. They are composed on the basis of relevant parameters agreed upon by the linguistic community and should include spoken and written, formal and informal language representing various social and situational strata. The idea behind reference corpora is that they can be used for a large variety of purposes, thus, rendering most special corpora unnecessary. They are also the point of reference when it comes to measuring the distortion of special corpora. They are used as benchmarks for lexicons and for the performance of generic tools and specific language technology applications. They are large in size; 50 million words is considered to be the absolute minimum; 100 million will become the European standard in a few years.

Monitor corpus. Language changes, and these changes should be reflected in a constant growth rate of corpora, leaving untouched the relative weight of its components (i.e., the balance) as defined by the parameters. The same composition schema should be followed year by year, the basis being a reference corpus with texts spoken or written in one single year.

Opportunistic corpus. The opportunistic corpus is an inexpensive alternative to the reference corpus. It is a collection of electronic texts that can be obtained, converted, and used free or at a very modest price; and its composition principle is that one should take all one can get and try to fill in blank spots as soon as they are recognized. Their place is in environments where size and corpus access do not pose a problem. The opportunistic corpus is a virtual corpus in the sense that the selection of an actual corpus (from the opportunistic corpus) is up to the needs of a particular project. Today's monitor corpora usually are opportunistic corpora.

Comparable corpus. For multilingual research and applications, corpora in each language are needed that follow the same composition pattern and, thus, can be used for language comparison. Opportunistic corpora cannot fulfill this claim. The focus is, therefore, on reference corpora. The Commission of the European Community is funding a project whose main goal is the creation of comparable reference corpora (of 50 million words each) for all the official languages of the European Union including Catalan and Irish. Comparable corpora are an indispensable source for bilingual and multilingual lexicons and a new generation of dictionaries. (10)

Parallel corpus. Texts in one language and their translations into other languages constitute parallel corpora. They are the source for the detection of translation equivalents and, thus, can play an important role in the development of multilingual lexicons. In order to do this, parallel corpora must be aligned at least sentence by sentence, preferably phrase by phrase. Their disadvantage is that the language of translations is distorted and does not contain the full range of vocabulary and syntax. To compensate for this deficiency, one can set up reciprocate parallel corpora, corpora containing authentic texts as well as translations in each of the languages involved. This allows double-checking translation equivalents. Only if a collocation or phraseologism occurs also in authentic texts, is it counted as an acceptable equivalent.

3. Standardization and Validation

Anyone who wants to build up a corpus has to make a number of decisions about how to encode the text. A text is more than a stream of words; it contains much more information. It may have one or more authors, a title, chapter headings, tables and figures, footnotes, foreign language citations, font shifts, paragraphs, and many other features. What should be retained? How should it be encoded? Should one aim at reconstruction of the source text complete with its specific layout? There are no strict rules. However we decide, we have to abide by the decision we made for the entire corpus. If we mark footnotes in text A, we should also mark them in exactly the same way in text B. There must be an explicit list of the features to be encoded, and there must be a set of unambiguous codes. Only a text collection that has been encoded in such a consistent way should be called a corpus.

Why do we need consistency in the creation of corpora? Corpora are raw data waiting for linguistic analysis. To extract information from corpora, we have to use software. The more complicated our query is, the more sophisti-

cated our software must be. If we want to extract complex phraseologisms consisting of all headlines of newspaper texts containing a certain phraseologism, we must make certain that headlines are marked consistently.

That corpora must be marked up consistently is self-evident. All large centers for language resources have developed or obtained corpus exploitation software demanding explicit coding of such linguistic and extra linguistic features. Within the limitations imposed by the software, the centers were free to chose the codes. This was adequate for a time when there was little exchange of corpora and software. However, when there was more demand for language resources and when a new generation of corpus software had to be developed able to handle corpora of 100 million words or more and still yield results in interactive access in a reasonable time, cooperation, exchange, and distribution became more and more important.

In recent years, the Text Encoding Initiative (TEI), an international project with strong North American and Western European participation, has developed standards and guidelines for the encoding of all sorts of texts (spoken and written) as corpora to be used as language resources. Likewise, standards and guidelines were developed for the set-up and exchange of lexical and terminological data. (11) The TEI recommendations are based on the SGML standard of the International Standards Organization (ISO). In the present form, they demand much extra work for the corpus compiler if they are to be followed step by step--probably more than most institutions are willing to invest. Therefore, European language resources projects like the Network of European Reference Corpora (NERC) and the first PAROLE project singled out subsets of the TEI standards that should be adopted by any new corpus (and lexicon) project. (12) ISO is preparing a number of SGML-based standards for terminology databases and the exchange of terminological data. (13)

For some years now the European Expert Group on Language Engineering Standards (EAGLES), an activity funded by the European Commission, have developed recommendations for the linguistic (and to some extent also extra-linguistic) annotation of corpora and the data categories used for lexicons. (14) For some application areas, there are only preliminary reports; in other fields, operational guidelines are already available.

Building up an infrastructure of language resources only makes sense if these resources are standardized. As we said before, this is the only way to ensure the reusability of resources for unlimited applications within the global NLP community. Standardization also helps in making resources comparable and in building links between resources. Only if monolingual lexicons follow the same architecture and use the same categories, is it possible to merge them into a multilingual lexicon. All alignment software for parallel corpora oper-

ates on the assumption that all relevant phenomena are encoded consistently. Standardization is a precondition of an operational language resources infrastructure. Its importance cannot be exaggerated.

On the other hand, it is not necessary to adopt these standards for internal use in institutions. Whoever has assembled a consistently encoded corpus and has developed proprietary software for corpus exploitation is under no obligation to change it to the TEI or EAGLES standards. These common international standards have to be used when it comes to data interchange. If we want to exchange or distribute our resources or our tools, we must convert them to be conformant or at least compatible with the narrow set of recommendations agreed upon by the NLP community. This task can be carried out by appropriate conversion software, some of which is offered by companies or is available under public domain.

What exactly should be standardized? It is a hot issue of discussion how far standardization should stretch. Basically there are two fields where standards and guidelines are offered for corpora. The one concerns the encoding of all relevant features of the text that we want to include into a corpus, data that are there, e.g., in the printed text of a novel. We usually can identify the author(s), find the titles, determine if there are chapters, paragraphs, and other layout features serving a purpose for the text. Features like these have to be encoded. But, what about full stops? These dots at the end of a sentence have to be preserved; however, do they also have to be disambiguated? On the surface, they do not differ from the little dot at the end of an abbreviation, e.g., viz. or etc. There are reasons to disambiguate them: we have to know where a sentence starts and ends. If we want to search for the co-occurrence of specific words forming a phraseologism it is useful to confine the search to sentences as linguistic units. If we want to align parallel text, the minimal parameter for alignment is the sentence, marked by the full stop and not to be confused with the little dot indicating that the previous string is an abbreviation. But if we decide for disambiguation, we add linguistic information to the text, using our linguistic knowledge and competence. Any kind of corpus annotation means adding even more and potentially more questionable linguistic information. Admittedly, there is a lot of agreement among experts whether a given word is a noun, verb, adjective, or something else. Yet, how should we tag the first element of language technology, the English equivalent to the German word Sprachtechnologie? Is it really a noun, or a noun used as a modifier, like an adjective, or is it just the first constituent of a compound?

If we add linguistic information of this kind to a corpus, then we might obliterate the very objective we wanted to achieve by using the corpus. Cor-

pora are used for obtaining new linguistic knowledge. But if we use traditional knowledge when it comes to the extraction of evidence from a corpus, we may never get to the point where we may want to discard old categories and replace them by new ones confirmed by corpus evidence. It is from looking at unannotated corpora that we have learned a lot about the fuzziness of the concept of lexical units and that we now see a continuum from a word segment like -euro- via single words, multiword units, collocations, to whole phaseologisms. Particularly in multilingual applications, it is obvious that the translation equivalents we are looking for are only very rarely single words and sometimes can be complete phrases or even sentences.

Any linguistic information added to a text or corpus tends to replicate any bias inherent in that category. Therefore, we must be careful that standardization of linguistic features originally not present in a text but added by hand or automatically does not obscure the evidence we expect to obtain from corpora. Normative so-called language-independent tagsets and categories for word class, morphosyntactic or even semantic information, whether in corpora or in lexicons, are useful only insofar as the can establish comparability between resources of different origin. If these tagsets are used for extracting evidence for linguistic phenomena from corpora, results have to be evaluated with great care.

The creation of language resources is expensive. The more effort that has been spent on the composition, documentation, and encoding of textual features on the annotation of added linguistic information, and on the comparability with related resources, the more valuable a corpus (or a lexicon) will be. To use it for just one specific purpose would be a waste--it should be reused as often as possible, covering a wide range of applications. As we said above, standardization is an absolutely necessary precondition if the resources are not only to be used internally within one single institution but are also to be made available to the NLP community. While in the past the exchange of resources and resource-oriented tools were based on bilateral agreement between provider and user, distribution centers today serve as platforms for the language resources market. They will be accepted only if the material they offer for distribution proves valuable to the clients, and therefore, it has to be fully documented and standardized. The first center was the Linguistic Data Consortium (LDC) at Pennsylvania University, which was created more than five years ago. Recently, the European Language Resources Association (ELRA) was founded for the (Western) European market. It will be complemented by national language resource centers. The Trans-European Language Resources Infrastructure (TELRI), a Concerted Action funded by the European Commission, is serving as a bridge between Central and Eastern Europe

(including the Commonwealth of Independent States) on the one side and Western Europe on the other. These distribution centers will pass on submitted resources to academic and industrial users and will collect license fees. It is no longer necessary to contact the institution where the resources were created in first place.

Such an arrangement works only if the resources are not only documented but validated as well. What does validation mean for written resources? For corpora, it means that the corpus has the size it claims, that it is composed the way it claims, that it is encoded the way it claims, that all features encoded can be used for retrieval, that annotations used conform to a given standard, and, above all, that the error rate for encoding and annotation does not exceed a certain level. Validation guarantees the client that he gets what he ordered and that he can rely on the resources to the extent stated by the validation certificate.

Validation has to be carried out on an unbiased and neutral basis, and this means not by the institution where the resources were created. Some of the validation features are language-independent, like the conformancy of the text representation with accepted rules and standards for encoding. But, some are language-specific, e.g., a claim that a lexicon really covers the core vocabulary of a given language, that the meanings assigned to a lemma have been gained from corpus evidence, and that the error rate of morphosyntactic tagging does not exceed a certain level. Therefore, it makes sense that validation for written resources should be carried out by institutions where there is sufficient competence for the language in question. On the other hand, to make validation reliable, all recognized validation centers should follow a validation procedure prescribed by a standard that is accepted by the NLP community. Some projects for the development of validation procedures are under preparation; the European Commission has included validation as a new work item in the new Telematics Programme.

Validation is important for generic resources designed for reusability. These resources will always be necessary for language technology. However, the more sophisticated applications are being designed, the more specific the language resources backing them will have to be. These resources will either have to be produced from scratch or by adding value to existing generic resources. In either case, specific resources of this kind will probably not be distributed by large centers, but produced by one institution and then be passed on directly to the user on a bilateral agreement. This trend will lead to a reassessment of the validation issue in the long run.

4. The Trans-European Language Resources Infrastructure (TELRI)

The boom in language industry has brought with it a growing demand for more and better monolingual and multilingual resources. In Europe, only a joint effort of existing focal language institutions could be expected to harmonize existing and designed standardized new resources in compliance with the needs of dictionary makers and developers of language technology applications.

The European Commission, realizing the central role of language engineering in the emergent information and communication technology market, has supported a number of relevant infrastructure activities, serving the needs of the three 'colleges' – speech (spoken language), terminology, and written resources. These projects helped to set up a common infrastructure for the countries of the European Union and the European Economic Area (formerly EFTA), and at the same time encouraged formation of national language resources networks. After years of preparation, the new PAROLE II project with partners from all European Union countries will produce a first generation of harmonized, comparable generic textual and lexical reusable resources, meeting the basic demands of language technology.

But Europe is larger than the European Union. All European countries must be given the opportunity to participate on an equal level in academic and industrial research and development. In the COPERNICUS Programme, the European Commission provided a framework of projects aiming at the integration of activities in Central and Eastern Europe with complementary ones in Western Europe. Several projects currently underway deal with various aspects of speech, terminology, and written resources. One of these projects dealing primarily with written resources is the Trans-European Language Resources Infrastructure (TELRI). (15) Since it aims at including as many partners in Central and Eastern Europe as possible, it is set up as a Concerted Action rather than as a project proper. Its partners are 22 focal language and language technology institutions in 17 countries, six Western European and 11 Central and Eastern European countries. The partners in the West are also cooperating in the PAROLE projects, thus linking closely TELRI activities with Western European developments. For the time being, there are no partners from former Yugoslavia (with the exception of Slovenia) or from the Commonwealth of Independent States. However, formalized links have been established with the leading institutions in Croatia, Serbia, and Russia; and these associated partners are participating in TELRI activities.

The Concerted Action TELRI has an initial duration of three years, beginning in early 1995, and is working on a budget of about half a million ECU.

It is not a research project: rather, its goal is to create a viable infrastructure in order to establish a permanent platform for industry, research institutes and universities, and to supply the NLP community with precompetitive or public domain monolingual and multilingual language resources. These resources are: corpora, machine readable dictionaries and lexicons, lexical data bases, and software tools for the creation, reuse, maintenance, valorization, and exploitation of linguistic data.

The activities of TELRI are organized in Working Groups for specific tasks. The collection, documentation, and dissemination of relevant information on language resources, providers and users, their potentials, and their needs is a basic activity. TELRI will promote the formation of national language resource networks, and TELRI partners will act as focal nodes. They will also design small scale joint ventures with private industry in order to foster cooperation between academic research and development. TELRI will pool and enhance existing service activities, providing resources, expertise, consulting and training facilities. The central platform will be annual seminars directed at the needs of small- and medium-sized enterprises. TELRI will engage in European and global standardization and validation activities and contribute to the harmonization of already existing resources. It is organizing joint research in the field of corpus-based multilingual lexicography and the use parallel aligned texts.

The main impact of TELRI will be to make the whole of Europe a strong competitor in the emerging market of language engineering and communication technology. Central and Eastern European language centers will participate in the creation of harmonized monolingual and multilingual resources making them ideal partners for joint ventures. Their huge potential in the creation of language data and the development of generic and specific language-related software will help to establish Europe as a leading force in language technology.

References

1. European Commission - DG XIII (ed). 1994. Telematics Applications Programme (1994–1998). Work Programme 15 December 1994. Luxembourg: Publications of the European Commission (1994), pp. i-x (further referred to as: EC - DG XIII (1994))
2. EC - DG XIII (1994), Appendix A5, p. 2
3. EC - DG XIII (1994), Appendix A5, p. 4
4. EC - DG XIII (1994), Appendix A5, p. 5

5. EC - DG XIII (1994), Appendix A5, p. 4 (my italics)
6. European Commission - DGs XII, III, XIII, XVII (ed). 1995. INCO COPERNICUS Information Package. Edition 1995/1996. Luxembourg: Publications of the European Commission (1995).
7. Teubert, W. 1996. "Parallel Corpora and Multilingual Lexicography". In: International Journal of Lexicography. (Forthcoming)
8. Teubert, W. 1992. "Zur Behandlung von Präpositionalattributen im Wörterbuch". In: Cahiers d'Études Germaniques. Beiträge zur Lexikologie und Lexikographie des Deutschen 23/1992, p. 119-135
9. Sinclair, J. and J. Ball, 1995. "Text Typology" (External Criteria). Draft Version. Electronic Document on the Pisa EAGLES ftp server, Birmingham
10. LE-PAROLE Consortium (ed). 1995. "LE-PAROLE Annex I" (Technical and Financial Annex). Annex to the project proposal submitted under the 1994-1998 TELEMATICS Applications Programme of the European Commission. Pisa/Paris: Consorzio Pisa Ricerche/GSI-Erli (1995), Chapter 3.1.2
11. Inquiries to: C. Michael Sperberg-McQueen, University of Illinois at Chicago, Academic Computing Center (M/C 135), 1940 W. Taylor, Rm. 124, Chicago, IL 600612-7352, USA, E-mail: u35395@uicvm.uic.edu
12. NERC Consortium (ed). 1993. Network of European Reference Corpora. Final Report. Birmingham/Leyden/Malaga/Mannheim/Paris/Pisa (1993); PAROLE (final report to be published in spring 1996; inquiries to Professor Antonio Zampolli, ILC Pisa, E-mail: eagles@ilc.pi.cnr.it
13. Inquiries to: Secretariat of ISO TC 37, c/o DIN Berlin, Burggrafenstraße 6, D-10787 Berlin, Germany
14. Information on EAGLES guidelines and recommendations can be obtained through the EAGLES ftp server at Pisa: nicolet.ilc.pi.cnr.it (anonymous login)
15. Information via WWW: http://www.ids-mannheim.de/telri/telri.html or E-mail: telri@ids-mannheim.de